Statistical Analysis of Local 500 hPa ECMWF Ensemble Forecasts

Seijo Kruizinga, Royal Netherlands Meteorological Institute

s.kruizinga@hccnet.nl

Summary.

In this paper some results from a statistical analysis of 500 hPa ECMWF ensemble forecasts for Western Europe are presented. The aspects analyzed were the climate of the models, the overall skill of the forecasts, the ranked histograms of the ensembles and the spread-skill relationship. The spread-skill analysis indicates that apart from under-dispersion (or over-dispersion) a random error in the ensemble mean is an important component in the statistical inconsistency of the ensembles.

1. Introduction.

In December 1992 ECMWF introduced the Ensemble Prediction System (Palmer et al. 1993). Since then this prediction system has been improved several times. A major improvement was introduced in December 1996. At that time the number of members of the ensemble was increased from 33 to 51. Several verification studies (*Palmer et al.*1997, *Lalaurette* 1999, *Richardson* 2000) have shown that the Ensemble Prediction System (EPS) has, in the early medium range, a better performance than the deterministic model of ECMWF. These studies were mainly based on verification tools like Brier Score, Ranked Probability Score, ROC area and rank histograms (*Talagrand et al.*, 1997). Except for the ranked histograms these scores are intended to monitor the overall skill of a forecasting system. In *Kruizinga* (1999) it has been shown that skill scores respond only marginally to errors in the ensemble, whereas consistency / reliability is strongly influenced by errors in the ensemble. However, different types of errors can lead to the same response in the consistency / reliability. This means that the interpretation of results on consistency is not straightforward. In this paper we will perform a statistical analysis on 4 years of EPS forecasts. We will compare the results with results from error-free simulated ensembles.

The layout of this paper is as follows. First we will describe in section 2 the dataset used in this verification study and in section 3 we will present the results of the statistical analysis. In section 4 we will draw some conclusions.

2. The dataset.

For this study we extracted the 500 hPa forecasts from the EPS including the control run (total ensemble size 51) and the 500 hPa forecasts and analyses from the deterministic (operational) model. From the forecasts and analyses valid at 00 UTC we computed on a daily basis three 500 hPa flow indices. These indices give a measure for respectively the Zonality, the Meridionality and the Cyclonality of the 500 hPa flow pattern over Western Europe and the Eastern Atlantic ocean. The derivation and characteristis of these flow indices are described in *Kruizinga*, 1979. So each day in the period studied, 1 January 1997 up to and inclusive 31 December 2000, is characterised by the three values of these flow indices. For a given index (e.g Zonality) and forecast lead time (e.g. +60, indicated as day 3) we constructed the following time series of 1461 steps each:

- The Zonality derived from the Operational analysis
- The Zonality derived from the Operational forecast
- The Zonality derived from the Control forecast
- The Zonality derived from the First member in the Ensemble.
- The Zonalities ordered in ascending order of the 51 ensemble members
- The Ensemble Mean, the usual mean over the 51 members.
- The Ensemble Spread, the standard deviation of the 51 members.

The total dataset consisted of 24 (8 (timesteps day 3 to 10) times 3 (indices)) of such sets of time series. In the analysis we studied the warm season and the cool season separately. In this paper we will only present the results for Zonality in the cool season. The results obtained for other flow indices and the warm season were highly similar.

3. Results

On the basis of the data described in the previous paragraph we studied 4 aspects of the forecasts.

3.1 Climate of the models.

In total the dataset contained 4 different deterministic forecasts for the observed value of the indices namely, Operational, Control, First Member and Ensemble Mean. In Fig. 1 (left panel) the bias (mean error) of each of the forecasts is plotted versus the lead time. The vertical scale in this figure (and next figures when appropriate) is rescaled with the standard deviation of the observations. So a bias of -0.05 means that the bias is only 5% of the standard deviation of the observations. Clearly the biases are relatively small. In the right panel of Fig.1 the standard deviations of the forecasts are plotted. As is seen the standard deviations of Operational, Control and First Member are close to 1.0 out to day 10. This indicates that the models show the same daily variation as the observations. The Ensemble Mean, however, tends to climatology at longer forecast ranges.

3.2 Average skill of the forecasts.

The average skill of the forecasts was determined with two different verification scores: the correlation between the forecasts and the observations and the standard deviation of the forecast error (forecast - observed). In Fig. 2 (left) the results for the correlation coefficient are shown. This figure shows that at longer ranges the Ensemble Mean is the best forecast. The First Member is clearly the worst forecast. Operational and Control are almost equal in skill. In Fig. 2 (right panel) the standard deviation of the forecast errors is plotted. These results confirm the results obtained with the correlation. Note that the standard deviation of the error is almost equal to the standard deviation of the observations at day 9 for Operational and Control. In this figure we also plotted the average of the daily Spreads. For statistically consistent ensembles the standard deviation of the forecast errors of the Ensemble Mean should be near to the average Spread (*Hou et al.*, 2001). However, Fig. 2 shows that the standard deviation is clearly higher than the mean Spread. Clearly the ensembles are not statistically consistent.

3.3 Consistency.

The ordered ensembles were used to construct the rank histograms. For each forecast ensemble the rank of the highest member below the observation was noted (observations below the lowest member were assigned rank 0). In Fig. 3 the frequency distributions of the ranks of the forecast ensembles for day 4 and 7 are plotted. For statistically consistent ensembles all the frequencies should be equal to 1/52. Clearly at day 7 but also at day 4 the frequencies of the end categories are higher than expected. This higher frequency at the end-categories has been observed in many other studies for a wide range of parameters and is usually interpreted as "the ensemble is under-dispersive" (*Mullen and Buizza*, 2000).

3.4 Spread-Skill relationship.

The main reason for performing Ensemble Predictions is the assumption that the predictability of the atmosphere varies from day to day and that the Spread of the ensemble reflects this predictability. The last part of this assumption was tested in the following way. The forecasts were categorized according to Spread: the forecasts with the 20% lowest Spreads were assigned to the class "Very good", the next 20% to the class "Good" and so on for "Normal", "Bad" and "Very bad". Within each class the standard deviation of the forecast errors and the average Spread was computed. For consistent ensembles with a good Spread-Skill relationship this standard deviation and average Spread should be more or less equal. In Fig. 4 the rescaled standard deviation in the classes "Very good","Good" etc. is plotted versus the rescaled average Spread, for day 4, 7 and 10 respectively. The dotted diagonal indicates the expected position of the plots. Clearly there is a relationship between Spread and standard deviation of forecast errors. However, for the classes "Very bad" the plots tend to be near or below the diagonal whereas for the "Very good" class the plots tend to lie above the diagonal. In Fig. 5 this can be seen more clearly. In this figure we plotted the ratio of "standard deviation" over Spread versus Spread. In this case we expect the plots near to 1.0, independent of the Spread.

The triangle plots in both figures represent the results obtained when the same Spread-Skill analysis was performed on data obtained with simulated (error-free) ensembles (*Kruizinga*, 1999). The characteristics of the simulated ensembles were set approximatedly equal to the characteristics of day 7 ensembles from EPS. Clearly the simulated ensembles behave in accordance with our expectation. However, the results obtained for the atmospheric ensembles cannot simply be attributed to under- or over-dispersion only. Clearly there is some additional error component that plays a more important role in high skill cases. A possible candidate for such an additional error component is a random error in the Ensemble Mean. This option was investigated by adding a random error to the means of the simulated ensembles. The results obtained with this type of simulated ensembles are shown in Fig. 6 for the Spread-Skill relationship and in Fig.7 for the rank histogram. The similarity between the simulated and the EPS ensembles is clear, indicating that a rank histogram with an overpopulation of the end categories can also be the result of an error in the ensemble mean and thus not necessarally indicates underdispersion.

4. Conclusions.

The results shown in the previous paragraph clearly show that:

- Single model run forecasts show the correct climate e.g. no bias and a standard deviation of the forecasts approximatedly equal to the standard deviation of the observations (Fig. 1)
- The Ensemble Mean tends to climatology at higher lead times (Fig. 1).
- The overall skill of the Ensemble Mean is higher than the skill of single run forecasts (Fig. 2)
- The Ensemble Spread underestimates the forecast error (Fig. 2).
- The observed values are observed too often outside the ensemble (Fig. 3)

Through a Spread-Skill analysis it was shown that:

- Spread and Skill are related and lower Spreads indicate higher Skill (Fig. 4).
- The underestimation of forecast errors is observed most clearly in cases of low Spread (or high skill) (Figs 4, 5)
- Through comparison with simulated ensembles it was shown that this effect is probably the result of random errors in the Ensemble Mean (Figs 6, 7)
- Such a random error in the Ensemble Mean explains the higher frequencies in the extreme categories of the rank histogram as well (Fig 7).

The origin of the random error in the Ensemble Mean can not be derived from this analysis. A possible candidate, however, is the random error in the initial analysis resulting in a random error in the Ensemble Mean.

Acknowledgement.

The support of Kees Kok in the preparation of this document is gratefully acknowledged.

References.

Hou, D., E. Kalnay and K.K. Droegemeier, 2001: Objective Verification of the SAMEX '98 Ensemble Forecasts, *Mon. Wea. Rev.*, **129**, 73-91.

Kruizinga, S., 1979: Objective classification of daily 500 mbar patterns, Sixth Conference on Probability and Statistics in atmospheric Sciences, Banff, Canada

Kruizinga, S., 1999: Verification of EPS forecasts. Seventh Workshop on Meteorological Operational Systems, Shinfield Park, Reading, UK, ECMWF, page 41-49

Lalaurette, F., 1999: Use of ECMWF products and performance of the forecasting system. Seventh Workshop on Meteorological Operationel Systems, Shinfield Park, Reading, UK, ECMWF, 25-37

Mullen, S.L. and R. Buizza, 2000: Quantitative Precipitation Forecasts over the United States by the ECMWF Ensemble Pediction System. *Mon. Wea. Rev.*, **129**,638-663

Palmer T.N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tibbia, 1993: Ensemble prediction. ECMWF Seminar Proc. on Validation of Models over Europe, **Vol. 1**, Shinfield Park,Reading, Uk, ECMWF, 21-66

Palmer T.N., R, Buizza, and F. Lalaurette, 1997: Performance of the ECMWF Ensemble Prediction System, Sixth Workshop on Meteorological Operational Systems, Shinfield Park, Reading, UK, ECMWF, page 19-30

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system, Quart. *J.Roy.Meteor.Soc.*, **126**, 649-668

Talagrand, O., R. Vautard and B. Strauss, 1997: Evaluation of probabilistic prediction systems. Workshop on Predictability, Shinfield Park, Reading, UK, ECMWF, pp 1-25.

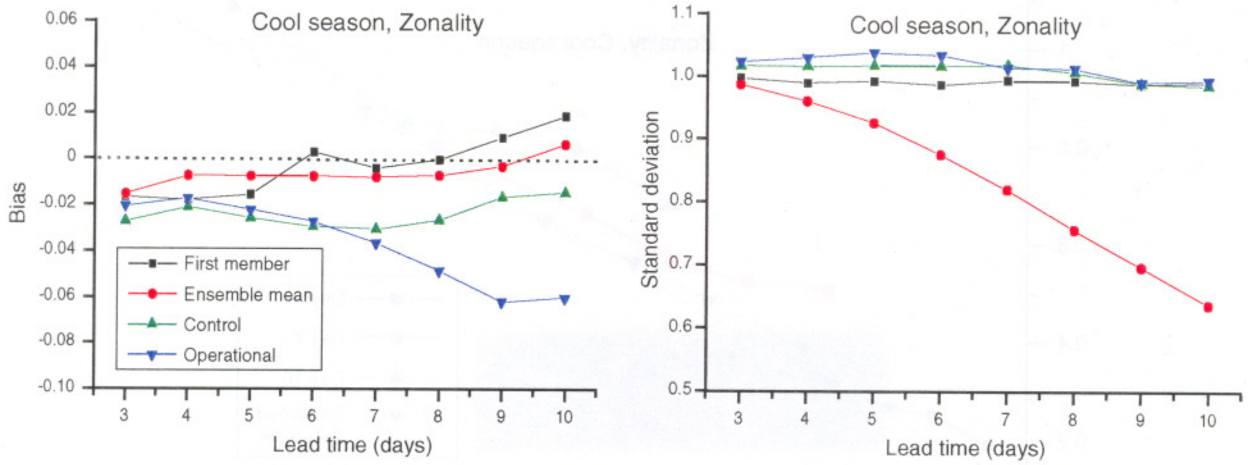

Fig. 1 Bias (left) and standard deviation (right) of the respective forecasts (vertical axis rescaled see text).

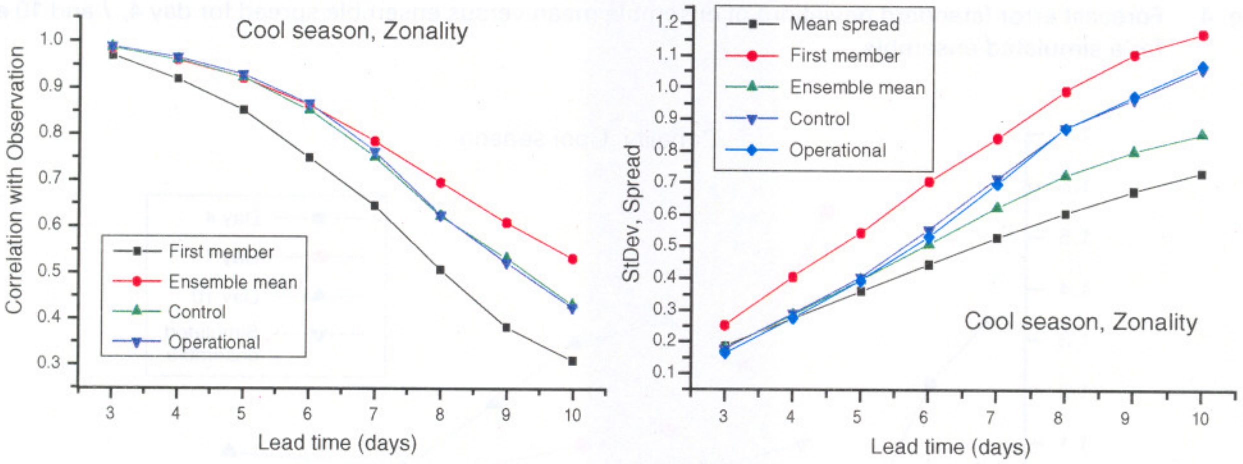

Fig. 2 Correlation between forecasts and observations (left) and standard deviation of forecast errors and mean ensemble spread (right).

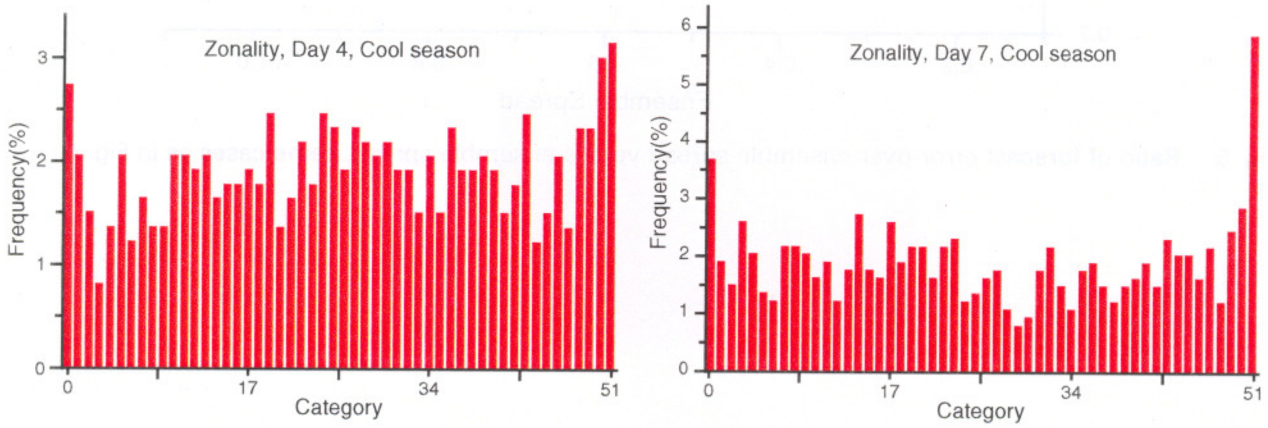

Fig. 3 Rank histograms for day 4 and day 7 respectively (note the different vertical scales)

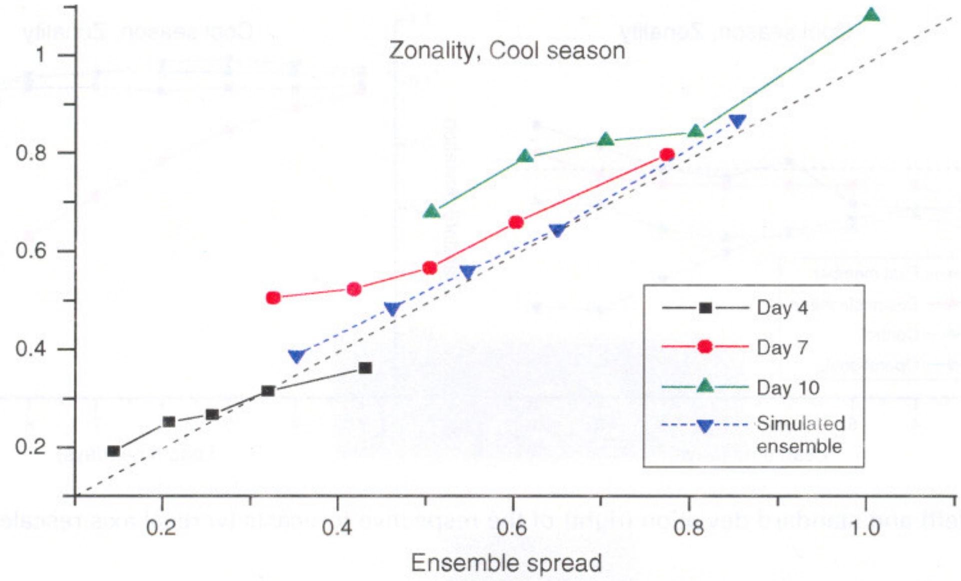

Fig. 4 Forecast error (standard deviation) of ensemble mean versus ensemble spread for day 4, 7 and 10 and for a simulated ensemble.

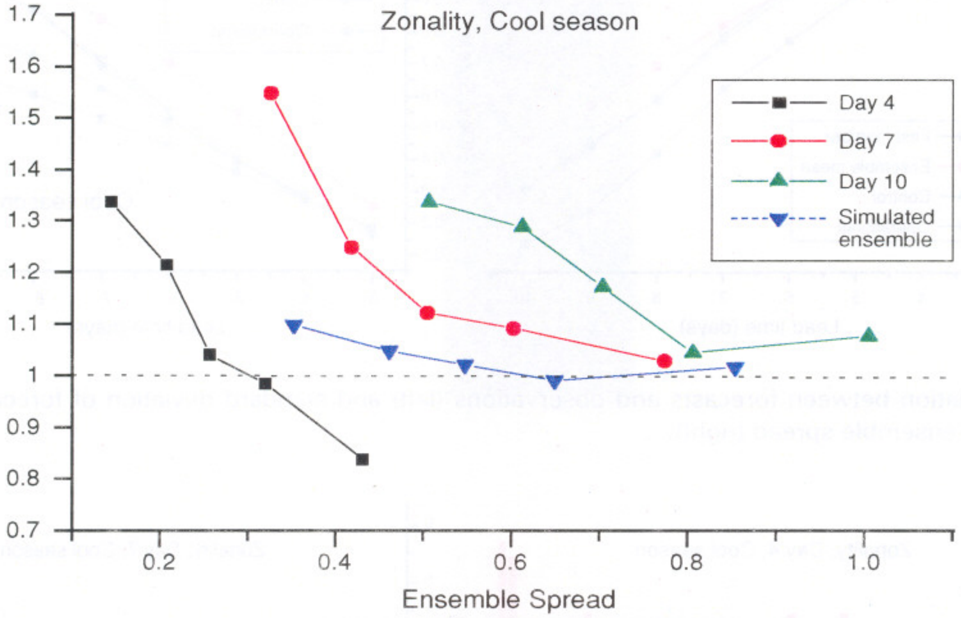

Fig. 5 Ratio of forecast error over ensemble spread versus ensemble spread, same cases as in Fig. 4.

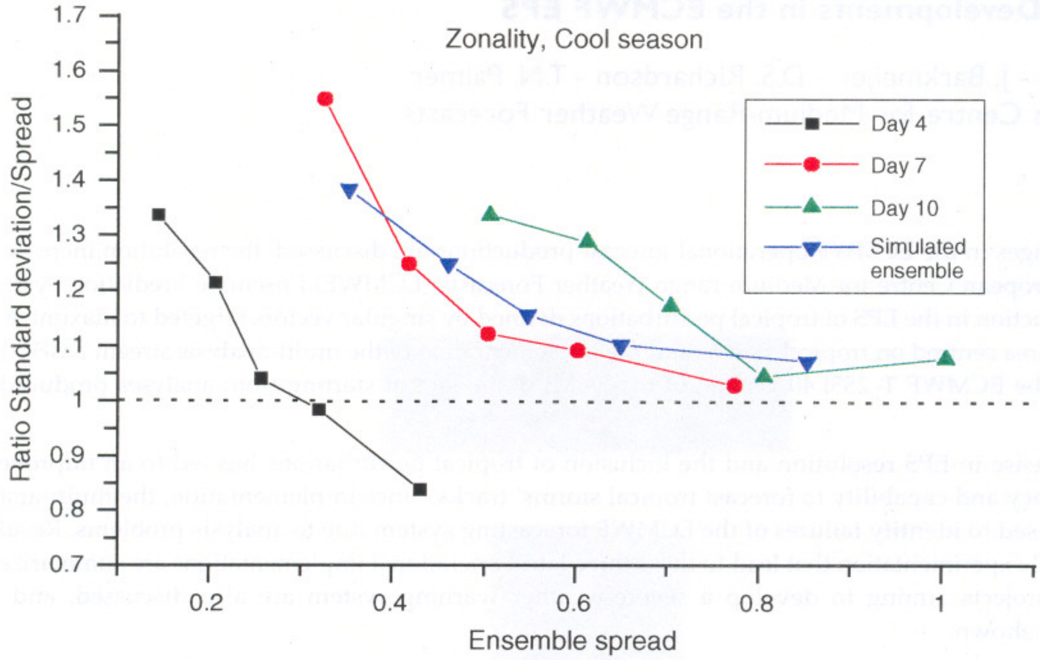

Fig. 6 As fig. 5 but now with a random error in the means of the simulated ensemble.

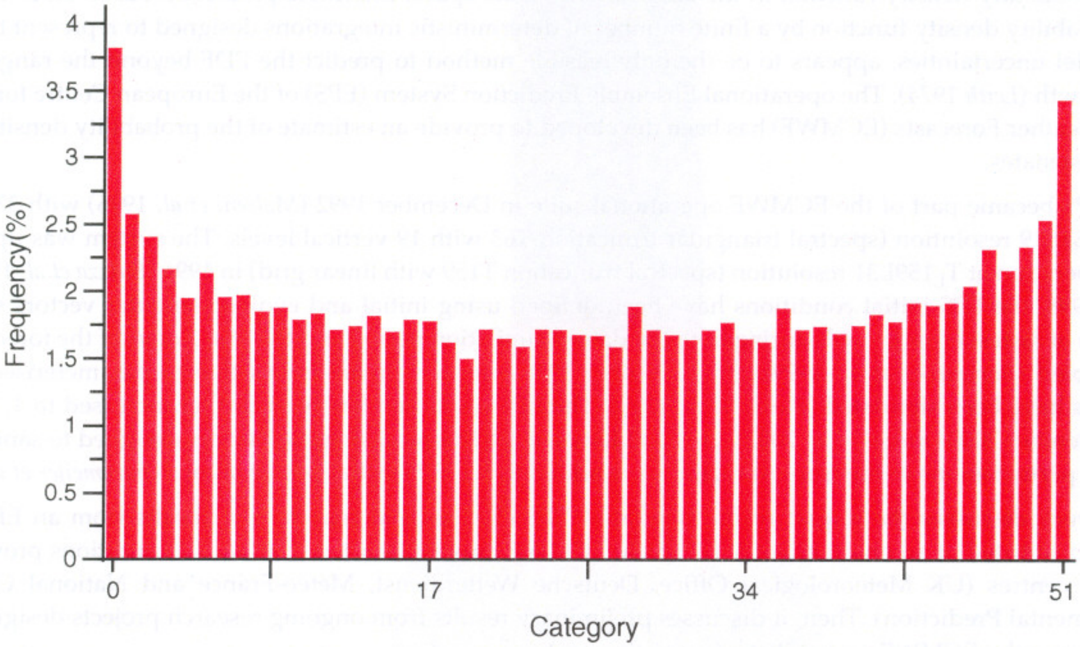

Fig. 7 Rank histogram of the simulated ensemble.